ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2015) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification ☆

Seyed Omid Sadjadi, John H.L. Hansen *

*Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75080-3021, USA*

Received 17 December 2013; received in revised form 20 April 2015; accepted 21 April 2015

**Abstract**

Adverse noisy conditions pose great challenges to automatic speech applications including speaker and language identification (SID and LID), where mel-frequency cepstral coefficients (MFCC) are the most commonly adopted acoustic features. Although systems trained using MFCCs provide competitive performance under matched conditions, it is well-known that such systems are susceptible to acoustic mismatch between training and test conditions due to noise and channel degradations. Motivated by this fact, this study proposes an alternative noise-robust acoustic feature front-end that is capable of capturing speaker identity as well as language structure/-content conveyed in the speech signal. Specifically, a feature extraction procedure inspired by the human auditory processing is proposed. The proposed feature is based on the Hilbert envelope of Gammatone filterbank outputs that represent the envelope of the auditory nerve response. The subband amplitude modulations, which are captured through smoothed Hilbert envelopes (a.k.a. temporal envelopes), carry useful acoustic information and have been shown to be robust to signal degradations. Effectiveness of the proposed front-end, which is entitled mean Hilbert envelope coefficients (MHEC), is evaluated in the context of SID and LID tasks using degraded speech material from the DARPA Robust Automatic Transcription of Speech (RATS) program. In addition, we investigate the impact of the dynamic range compression stage in the MHEC feature extraction process on performance using logarithmic and power-law non-linearities. Experimental results indicate that: (i) the MHEC feature is highly effective and performs favorably compared to other conventional and state-of-the-art front-ends, and (ii) the power-law non-linearity consistently yields the best performance across different conditions for both SID and LID tasks.

*Keywords:* Language identification; MHEC; Mismatch conditions; Robust features; Speaker identification

## 1. Introduction

Research in fields of speaker and language identification (SID and LID) has seen significant advancements in recent years. Current state-of-the-art acoustic SID and LID systems are primarily focused on channel and session mismatch compensation techniques in the back-end. The research trend in this domain has gradually migrated from joint factor analysis (JFA) based methods, which attempt to model the speaker/language and channel subspaces separately (Kenny et al., 2007), towards the i-vector approach that models both speaker and channel into a single space

☆ This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch. The authors would like to thank Mitchell McLaren and Aaron Lawson of SRI for providing development lists for speaker and language identification experiments.

* Corresponding author.
*E-mail address:* john.hansen@utdallas.edu (J.H.L. Hansen).
*URL:* http://crss.utdallas.edu (J.H.L. Hansen).

http://dx.doi.org/10.1016/j.specom.2015.04.005
0167-6393/© 2015 Published by Elsevier B.V.

termed the total variability subspace (Dehak et al., 2011). Various classifiers, models, and scoring methods are conveniently applied to i-vectors. These include support vector machines (SVM) (Dehak et al., 2011; Han et al., 2013), probabilistic linear discriminant Analysis (PLDA) (Prince and Elder, 2007; Garcia-Romero and Espy-Wilson, 2011; McLaren et al., 2013), adaptive Gaussian back-end (AGB) (Lawson et al., 2013), neural networks (NN) (Lawson et al., 2013; Matějka et al., 2012), and the simple yet effective cosine distance (CD) based scoring (Dehak et al., 2011) which is typically combined with LDA (Fukunaga, 1990) followed by within-class covariance normalization (WCCN) (Hatch et al., 2006).

In spite of progress seen in back-end advancements, several research efforts have been made recently that aim at development of robust front-end solutions to alleviate the adverse impact of acoustic mismatch on performance. These include robust speech activity detection (SAD) (Sadjadi and Hansen, 2013a,b; Kinnunen and Rajan, 2013), speech enhancement for noise reduction or de-reverberation (Sadjadi and Hansen, 2010, 2012; Matějka et al., 2012; Han et al., 2013; Godin et al., 2013), robust and low variance spectrum estimation (Hanilci et al., 2012a,b; Kinnunen et al., 2012; Alam et al., 2013), as well as blind and model based feature normalization techniques (Hasan et al., 2013; Castaldo et al., 2007). Another line of research in the front-end domain has focused on the design of alternative acoustic features which are not only capable of capturing the speaker/language identity conveyed in the speech signal, but also robust to environmental and channel distortions (e.g., see Shao et al. (2007), Li and Huang (2010), Sadjadi and Hansen (2011), Ganapathy et al. (2012), Lawson et al. (2013)).

Although originally designed to represent acoustic spaces of different phonemes for ASR, mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) have been the most widely used features for speaker and language recognition tasks, probably because they provide acceptable performance under matched conditions. In addition, this makes it possible to easily integrate SID/LID and ASR systems. However, it is well-known that MFCC based systems are susceptible to acoustic mismatch scenarios that occur between training and test conditions due to background noise, room reverberation, communication channel, and speaker variability (e.g., vocal effort), just to mention a few. There are several factors contributing to this susceptibility, among which the following are most dominant: (1) the periodogram based spectrum estimation approach in standard MFCC extraction is not robust to noise and channel distortions (Kinnunen et al., 2012), (2) the auditory model (i.e., the mel scale) used in MFCC may not necessarily be optimal for speaker/language recognition (Zhou et al., 2011; Kinnunen et al., 2013), and (3) the logarithmic nonlinearity used in MFCC to compress the dynamic range of filterbank energies is not immune to noise and channel distortions (Sarikaya and Hansen, 2001; Ravindran et al., 2006). These issues have been the primary motivation in the design of alternative robust acoustic features that are less affected by environmental conditions such as background noise and/or transmission channel.

In this study, we propose an alternative acoustic feature representation that is shown to bring robustness to speaker and language identification systems under noisy and reverberant mismatched conditions (Sadjadi and Hansen, 2011, 2013a, 2014; Lawson et al., 2013). The feature extraction procedure, which is inspired by the human auditory processing, uses the Hilbert envelope of Gammatone filterbank outputs to compute subband amplitude modulations (AM). Slowly varying AMs in narrow frequency bands carry useful acoustic information (Drullman et al., 1994; Assmann and Summerfield, 2004) and have proven to be robust to signal degradations in different speech applications such as ASR, SID and LID (Maragos et al., 1993; Kingsbury et al., 1998; Lawson et al., 2013; Mitra et al., 2013).

The effectiveness of the proposed acoustic feature front-end, which is entitled the mean Hilbert envelope coefficients (MHEC), is evaluated in the context of speaker and language identification tasks under actual noisy channel conditions using speech data from the DARPA program Robust Automatic Transcription of Speech (RATS). The RATS data, which is distributed by the Linguistic Data Consortium (LDC) (Walker and Strassel, 2012), consists of conversational telephone speech (CTS) recordings that have been retransmitted (through LDC's Multi Radio-Link Channel Collection System) and captured over 8 extremely degraded communication channels, labeled A–H, with distinct noise characteristics. The distortion type seen in RATS data is nonlinear and the noise is to some extent correlated with speech. We conduct our SID experiments with a state-of-the-art i-vector based system (Dehak et al., 2011) and report equal error-rate (EER) and false-alarm rate at 10% false-rejection rate (FA10m) as performance measures. As for LID experiments, a JFA based system (Kenny et al., 2007) without Eigenvoices is used to model the languages (this is referred to as an Eigenchannel system), and EER as well as average cost (Cavg) (NIST, 2009) are reported as performance metrics. We benchmark the MHEC feature against the conventional MFCCs as well as the noise-robust power normalized cepstral coefficients (PNCC) (Kim and Stern, 2012) and perceptually motivated minimum variance distortion-less response (PMVDR) features (Yapanel and Hansen, 2008a). Additionally, we investigate the impact of the dynamic range compression stage in the MHEC feature on SID and LID performance using logarithmic and power-law nonlinearities. It has been previously shown that, for ASR tasks, the root cepstrum is more immune to acoustic mismatch due to ambient noise compared to the logarithmic cepstrum (Alexandre and Lockwood, 1993; Sarikaya and Hansen, 2001). To the best of our knowledge, this study is the first to undertake comparison of root versus log-cepstrum for noise-robustness in SID and LID tasks.

It is worth noting here that in our earlier work, e.g., in Sadjadi and Hansen (2011), we only provide the procedure for the extraction of the MHEC feature without much detail regarding the various components involved in the extraction pipeline including the auditory inspired gammatone filterbank, the Hilbert operator and envelope, as well as the compression operator. Additionally, experimental results were only provided on rather clean data, e.g., on NIST SRE2010 (Sadjadi et al., 2012). Furthermore, the effectiveness of the MHEC for language recognition was not discussed, neither did we provide experimental results to support the robustness of the MHEC feature for real noisy tasks. Finally, performance comparisons were only provided against the MFCC feature in the past. The current manuscript, on the other hand, not only discusses the extraction process for the MHEC feature with far more details, but also presents results on real noisy data (with linear and nonlinear additive and channel noise) for both SID and LID tasks. Performance comparisons are also provided with respect to several state-of-the-art feature representations that have recently shown promise on similar real noise tasks.

## 2. Mean Hilbert envelope coefficients: MHEC

In this section, the procedure for extraction of the acoustic feature parameters based on the Hilbert envelope of Gammatone filterbank outputs, is described. A block diagram illustrating the proposed feature extraction scheme is depicted in Fig. 1.

First, the preemphasized speech signal $s(t)$ is decomposed into 32 bands through a 32-channel Gammatone filterbank (Patterson et al., 1992). The filterbank center frequencies are uniformly spaced on an equivalent rectangular bandwidth (ERB) scale between 200 and 3400 Hz (assuming a telephone bandwidth at a sampling rate of $F_s = 8$ kHz). The ERB for channel $j = 1, 2, \ldots, 32$ is computed as,

$$\mathrm{ERB}_j = \frac{f_j}{Q_{ear}} + B_{min}, \qquad (1)$$

where $Q_{ear} = 9.26449$ and $B_{min} = 24.7$ are known as Glasberg and Moore parameters (Glasberg and Moore, 1990), and $f_j$ is the center frequency in Hertz. The frequency response of the 32-channel Gammatone filterbank is illustrated in Fig. 2. As seen from the figure, the filterbank essentially consists of a set of bandpass filters whose impulse responses are the product of a gamma function and a tone that represent the frequency response associated with a particular point on the basilar membrane of the cochlea. A Gammatone filter is defined by its time-domain impulse response which has a closed-form as (Patterson et al., 1992),

$$h(t,j) = \delta t^{\tau-1} \exp\left(-2\pi b(f_j)t\right) \cos\left(2\pi f_j t + \theta\right), \qquad (2)$$

where $\delta$ and $\tau$ denote the magnitude of the response and the filter order, respectively, $b(f_j)$ is the filter bandwidth with $f_j$ being the center frequency in Hz, and $\theta$ is the initial phase. The output signal at each channel is obtained as the convolution of the speech signal $s(t)$ with the impulse response at that channel, i.e.,

$$s(t,j) = s(t) \ast h(t,j). \qquad (3)$$

The subband signal $s(t,j)$ is known to have both an amplitude-modulation (AM) and a frequency-modulation (FM) structure as follows (Maragos et al., 1993),

$$s(t,j) = a(t,j) \cdot \cos\left[\phi(t,j)\right], \qquad (4)$$

where $a(t,j)$ and $\phi(t,j)$ represent instantaneous amplitude and phase signals at the $j$th channel, respectively. This is illustrated in Fig. 3 for a sample subband speech signal at a center frequency of $f_j \approx 1000$ Hz, where the slowly varying instantaneous envelope (shown in dashed red) is

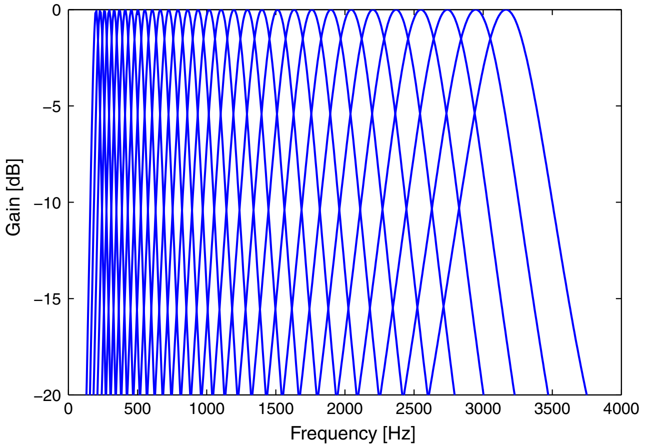

Fig. 2. Frequency response of the 32-channel Gammatone filterbank. Center frequencies are uniformly spaced on ERB scale in [200 3400] Hz range. A sample-rate of 8000 Hz is assumed.

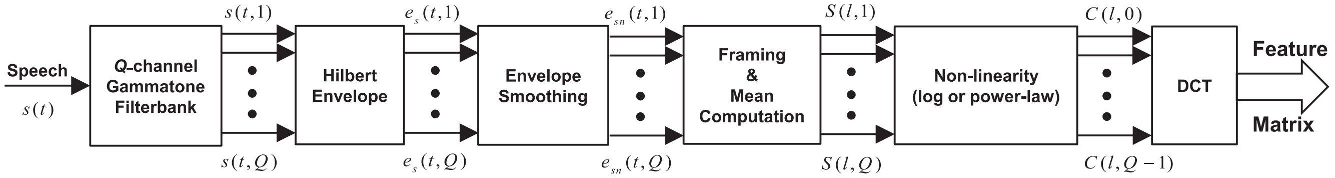

Fig. 1. Block diagram of the proposed feature extraction scheme. The symbols represent the output signals at each stage.

Please cite this article in press as: Sadjadi, S.O., Hansen, J.H.L., Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.04.005

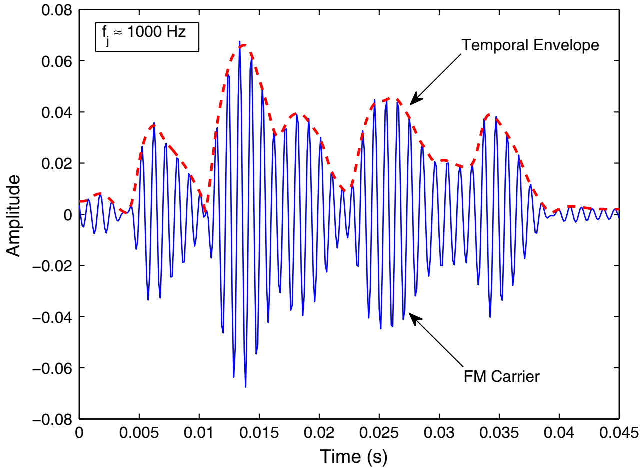

Fig. 3. A sample subband speech signal at a center frequency of $f_j \approx 1000$ Hz. The instantaneous temporal envelope is shown in red (dashed). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

superimposed over the FM carrier. The temporal envelope of speech is also known as the message which fluctuates at the rate of movements in human speech production apparatus including jaws, tongue, and lips (excluding the vocal folds). The instantaneous frequency of the carrier signal is a function of the vibration speed of vocal folds.

Next, since we are primarily interested in slowly varying amplitude modulations $|a(t,j)|$ rather than the fine structure (i.e., instantaneous frequency), $\dot{\phi}(t)$, the temporal envelope of the $j$th channel output $s(t,j)$ is computed as the squared magnitude of the analytical signal obtained using the Hilbert transform. This approach is known as the Hilbert transform demodulation (HTD) which is employed to separate the AM signal from its FM counterpart. More specifically, let

$$s_a(t,j) = s(t,j) + i\hat{s}(t,j), \tag{5}$$

denote the analytical signal, where $\hat{s}(t,j)$ is the Hilbert transform of $s(t,j)$, and $i$ is the imaginary unit. The theoretical and computational advantages of the Hilbert transform over other alternative demodulation techniques such as the Teager–Kaiser energy operator (TEO) (Maragos et al., 1993) have been discussed in Vakman (1994), which describes the problem of AM–FM decomposition from the perspective that AM and FM components of a signal can be derived from the sum of the real signal and its conjugate. Vakman (1994) defines three physical conditions for the conjugation operator: (1) continuity, (2) homogeneity (i.e., invariance to scaling), and (3) harmonic correspondence (i.e., a single frequency tone must retain its constant amplitude and frequency after the operation, with only a 90-degree shift in phase). It was shown that the Hilbert operator is the only transform that satisfies all three condition, while the TEO violates the continuity condition. Accordingly, when compared to the TEO, the Hilbert operator is more accurate for noisy signals, and enables elimination of distortion in special cases with wideband noise (Schimmel, 2007), which is a common case in channel distorted RATS data. Finally, Vakman (1994) showed that the Hilbert transform is easier and faster to carry out regarding the computer processing.

The temporal envelope $e_s(t,j)$ is thus calculated as,

$$e_s(t,j) = s^2(t,j) + \hat{s}^2(t,j), \tag{6}$$

with,

$$|a(t,j)|^2 \approx e_s(t,j). \tag{7}$$

Here, $e_s(t,j)$ is also called the Hilbert envelope of the signal $s(t,j)$. At the next stage, in order to further suppress the remaining redundant high frequency components, thereby reducing the approximation error in (7), the Hilbert envelope $e_s(t,j)$ is smoothed using a low-pass filter with a cut-off frequency of $f_c = 20$ Hz as,

$$e_{sn}(t,j) = (1-\eta)e_s(t,j) + \eta e_{sn}(t-1,j), \tag{8}$$

where the subscript $n$ denotes the smoothed (or normalized) envelope, and $\eta$ is a smoothing factor (inversely) exponentially proportional to the cut-off frequency as,

$$\eta = \exp\left(\frac{-2\pi f_c}{F_s}\right). \tag{9}$$

As noted earlier, the smoothed subband Hilbert envelopes which represent the envelope of the auditory nerve response, carry useful acoustic information (Drullman et al., 1994; Assmann and Summerfield, 2004) and have been shown to be robust to signal degradations due to background noise and communication channel for different speech applications (Maragos et al., 1993; Kingsbury et al., 1998; Lawson et al., 2013; Mitra et al., 2013). Additionally, it is well-known that the subband AM signals at center frequencies above 1 kHz fluctuate at fundamental frequency (i.e., pitch) of speech (Weintraub, 1985; Hu and Wang, 2004). Our hypothesis is that this property can help as an acoustic cue to better discriminate among different speakers as well as languages.

At the next stage, the smoothed Hilbert envelope $e_{sn}(t,j)$ is blocked into frames of 25 ms duration with a skip rate of 10 ms. A Hamming window is applied to each frame to minimize discontinuities at the edges as well as the correlation between adjacent frames. To estimate the temporal envelope amplitude in frame $l$, the sample means are computed as,

$$S(l,j) = \frac{1}{M}\sum_{t=0}^{M-1} w(t) \cdot e_{sn}(t,j), \tag{10}$$

where $w(t)$ denotes the Hamming window and $M$ is the frame size in samples. Note that $S(l,j)$ is also a measure of the spectral modulation energy at the center frequency of the $j$th channel, and therefore provides a short-term spectral representation of the speech signal $s(t)$.

To compress the dynamic range of the estimated spectral parameters, $S(l,j)$, a nonlinear operator such as the natural logarithm (Davis and Mermelstein, 1980) or root (Hermansky, 1990) is commonly applied in the extraction of cepstral features. It has been known in the ASR community that the logarithmic spectrum compression adopted in conventional cepstral analysis is susceptible to noise (Alexandre and Lockwood, 1993; Sarikaya and Hansen, 2001; Ravindran et al., 2006; Kim and Stern, 2012). As a remedy, the notion of root-cepstrum (as opposed to log-cepstrum) was first introduced in Lim (1979) for homomorphic deconvolution, and later was extended to acoustic features in Hermansky (1990), Alexandre and Lockwood (1993), Sarikaya and Hansen (2001) to achieve noise robustness in ASR tasks. Inspired by such observations from ASR experiments, in this study, we evaluate the effectiveness of the root-cepstrum versus log-cepstrum for SID and LID tasks. In particular, the spectral parameters $S(l,j)$ are either compressed using the natural logarithm or power-law nonlinearity with an exponent term $\gamma$ as,

$$C(l,j) = [S(l,j)]^{\gamma}, \tag{11}$$

where $\gamma$ is set to 1/15 as suggested in Kim and Stern (2012). The motivation behind employing a nonlinear function as in (11), is twofold. First, the root compression with an exponent of $\gamma = 1/15$ provides a better fit to the psychophysical data observed from subjective experiments (i.e., involving human listeners), particularly for sound pressure levels below 0 dB SPL (Kim and Stern, 2012). In other words, it approximates more accurately the nonlinear relation between the input sound intensity and auditory nerve firing rate. Second, for frequency channels with energies much smaller than one, that is $S(l,j) \ll 1$, the compressed output has a value close to zero as opposed to negative infinity. Note that the logarithmic compression artificially increases the variance for such spectral components, resulting in a greater variability of estimated cepstral features which has been shown to be a major source of performance loss for speech applications.

After the nonlinearity stage, which uses either the log or root compression, the discrete cosine transform (DCT) is applied as,

$$c(l,q) = \sum_{j=0}^{Q-1} C(l,j) \cdot \cos\left[\frac{\pi q}{2N}(2j+1)\right], \quad q = 0,\ldots,31. \tag{12}$$

The DCT is used to: (1) convert the spectrum to the cepstrum, and (2) decorrelate the various feature dimensions because, as shown in Fig. 2, there is significant overlap between adjacent filters. The latter is important because Gaussian mixture models (GMM) with *diagonal* covariance matrices of reasonable size can then be used to model the acoustic spaces of speakers (as opposed to *full* covariance matrices). A potential shortfall in using the DCT, which becomes apparent after a careful examination of (12), is that it spreads any local distortion (due to background noise and/or transmission channel) across all resulting cepstral coefficients. This becomes even more problematic if the nonlinearity stage introduces undesirable and artificial variabilities as well (e.g., very large negative values for very small spectral energies after the log compression). This constitutes another important reason for adopting a nonlinear operator that is less susceptible to processing and environmental artifacts.

The output of the DCT stage is a matrix of 32-dimensional cepstral features $c(l,q)$, entitled the mean Hilbert envelope coefficients (MHEC). For our SID experiments, only the first 20 coefficients (including $c_0$) are retained after the DCT. Because the cepstral representation of the speech spectrum is only a measure of the local pattern of the signal at a given frame, the first and second temporal cepstral derivatives are also computed over a 5-frame window and appended to the static features to capture the dynamic pattern of speech over time. This results in 60-dimensional feature vectors. In our LID experiments, on the other hand, only the first 7 coefficients (again including $c_0$) are kept after the DCT. Here, in order to capture the temporal context information, shifted delta cepstral (SDC) features are computed using the standard configuration parameters 7-1-3-7 ($N$–$d$–$P$–$k$) (Bielefeld, 1994; Torres-Carrasquillo et al., 2002). The SDC features are appended to the static cepstral coefficients resulting in a 56-dimensional feature vector.

## 3. Experiments

In this section, we provide brief descriptions for alternative noise-robust acoustic features considered in this study for comparison purposes. In addition, we present experimental setups for SID and LID tasks, i.e., speech data used as well as system descriptions.

### *3.1. Alternative robust acoustic features*

In addition to the conventional MFCCs, in this study, we consider evaluation of two other robust acoustic features for SID and LID tasks on noisy channel degraded data from the RATS program. The features are termed Perceptual MVDR cepstral coefficients (PMVDR) and Power Normalized Cepstral Coefficients (PNCC). We benchmark SID and LID performance with these features against that obtained with the MHEC feature. The extraction procedure for the PMVDR and PNCC features are briefly described in the following two subsections. It is worth noting that we utilize the same configuration parameters as in the MHEC to extract MFCCs, i.e., features are computed for 25 ms frames every 10 ms using a 32-channel mel filterbank spanning the frequency range [200 3400] Hz. We employ HCopy tool from the Hidden Markov model (HMM) toolkit (HTK) (Young et al., 2009) for MFCC extraction.

#### *3.1.1. Perceptual MVDR cepstral coefficients (PMVDR)*

The PMVDR front-end which uses a noise-robust perceptual linear prediction (LP) based spectrum estimation technique with minimum variance was proposed by Yapanel and Hansen (2008a). The acoustic features extracted using the perceptual Minimum Variance Distortionless Response (MVDR) spectrum have been shown to outperform the conventional MFCCs under noisy conditions for ASR (Yapanel and Hansen, 2008a) as well as speaker recognition applications (Lawson et al., 2011; Godin et al., 2013). The schematic diagram of the PMVDR front-end is depicted in Fig. 4. The processing steps for the PMVDR front-end are as follows:

The preemphasized speech signal is first transformed into the short-time Fourier transform (STFT) domain using a window length of 25 ms with a skip rate of 10 ms. Next, the power spectrum is calculated and directly warped towards a perceptual scale through a first order all-pass filter. The phase response of the first order filter is given as,

$$\hat{\omega} = \tan^{-1}\frac{(1-\alpha^2)\sin(\omega)}{(1+\alpha^2)\cos(\omega)+2\alpha}, \tag{13}$$

where $\omega$ and $\hat{\omega}$ represent the linear and warped frequencies, respectively. The parameter $\alpha$, which controls the degree of warping, is set to 0.31 to approximate the mel scale (assuming a sampling rate of $F_s = 8$ kHz). More information on the implementation of the direct warping can be found in Yapanel and Hansen (2008a). The warping term $\alpha$ in PMVDR can also be adjusted in a manner similar to vocal tract length normalization (Yapanel and Hansen, 2008b).

After obtaining the perceptually warped power spectrum for each frame, the perceptual autocorrelation lags are computed via the inverse FFT. Next, a $p$th order linear prediction (LP) analysis employing Levinson–Durbin recursion is performed. This is followed by calculating the $p$th order MVDR spectrum from the LP coefficients. The MVDR spectrum accurately models the peaks in the speech spectrum by successfully connecting the spectral peaks to form an upper spectral envelope which is expected to be more robust in noisy environments.

Finally, the PMVDR cepstral coefficients are obtained by taking the FFT of the MVDR spectrum, applying the natural logarithm, and taking the inverse FFT. Here, similar to the MHEC front-end, the first 20 cepstral coefficients are retained after IFFT (including $c_0$) and appended with the first and second temporal derivatives to form 60-dimensional feature vectors for SID experiments. For LID experiments, we append the SDC features to the first 7 cepstral coefficients (including $c_0$), resulting in a 56-dimensional feature vector for each frame.

### 3.1.2. Power normalized cepstral coefficients (PNCC)

The PNCC front-end, which is also motivated by the principles of human auditory system, includes several built-in nonlinear processing modules as well as a power-law (as opposed to logarithmic) compression stage to suppress additive noise and reverberation (Kim and Stern, 2012). It was originally proposed for robust ASR across a range of noisy conditions (Kim and Stern, 2012), and later adopted in several other studies for robust SID and LID applications as well (McLaren et al., 2013; Lawson et al., 2013). The block diagram of the PNCC front-end is shown in Fig. 5. The processing steps for the PNCC front-end are as follows:

The preemphasized speech signal is first transformed into the STFT domain using a Hamming window of 25 ms duration with a skip rate of 10 ms. Next, the power spectrum is calculated and warped on an ERB scale by weighting the magnitude-squared STFT outputs with the frequency response of a 32-channel Gammatone filterbank as shown in Fig. 2 (note the difference between time domain filterbank analysis in the MHEC feature versus the frequency domain analysis in the PNCC front-end). After obtaining the perceptually warped power spectrum for each frame, two distinct noise reduction algorithms are applied, namely Asymmetric Noise Suppression (ANS) and temporal masking. The ANS can be considered a spectral subtraction based noise suppressor (Boll, 1979) that is realized in each frequency band through a time-varying nonlinear filter. The filter tracks the noise as the lower envelope of the subband signals in a manner similar to the minimum statistics noise estimation proposed by

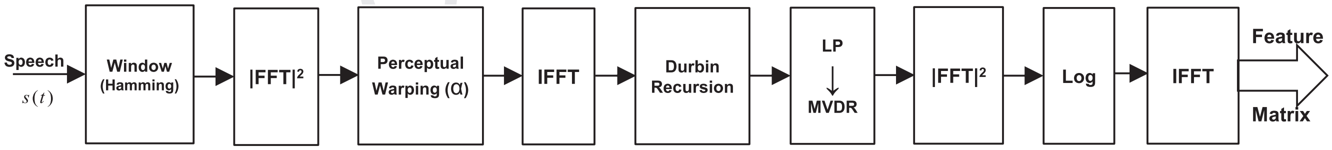

Fig. 4. Schematic block diagram of the perceptual MVDR (PMVDR) front-end.

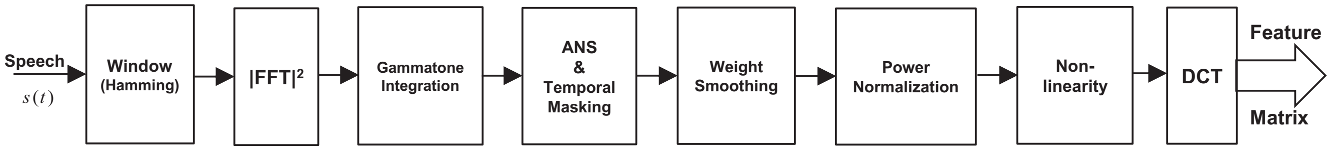

Fig. 5. Schematic block diagram of the PNCC front-end.

Please cite this article in press as: Sadjadi, S.O., Hansen, J.H.L., Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.04.005

Martin (2001). The temporal masking, on the other hand, enhances the envelope onsets within each subband via a nonlinear and single-pole filter. It is motivated based on the process observed in the human auditory system that appears to respond more strongly to the onset of an incoming power envelope than the trailing edge of the same envelope.

Following the noise compensation stage, the power spectrum response is smoothed in each band and normalized using a running average filter that compensates for a constant scale factor. This is important because in the next stage a power-law nonlinearity as in (11) is applied, therefore the scaling effect cannot be compensated for through cepstral mean subtraction (CMS). After the spectral components are root-compressed, DCT is applied to obtain the PNCC feature vectors. As in other front-ends considered in this study, temporal context modeling is applied to PNCCs to form 60-dimensional and 56-dimensional feature vectors for SID and LID experiments, respectively.

For the sake of clarity, configuration parameters used for the MHEC as well as other front-ends considered in this study are summarized Table 1.

### 3.2. Data

We evaluate the MHEC along with MFCC, PMVDR, and PNCC features in the context of SID and LID tasks using *actual* noisy and channel degraded speech material from the DARPA RATS program. The speech data is distributed by LDC (Walker and Strassel, 2012) and consists of conversational telephone speech (CTS) recordings that have been retransmitted through LDC's Multi Radio-Link Channel Collection System and captured over 8 extremely degraded communication channels, labeled A–H, with distinct noise characteristics. The type of distortion seen in RATS data is nonlinear (e.g., akin to clipping) and the noise is to some extent correlated with speech.

#### 3.2.1. SID task

A total of 8859 speech recordings, including clean source files along with their degraded (retransmitted) versions (channels A through H), from the data release LDC2012E63 for the RATS SID task (RATS, 2013), are partitioned into development, enrollment, and test sets. There are five languages spoken in the audio data: Levantine Arabic (alv), Dari (prs), Farsi (fas), Pashto (pus), and Urdu (urd). The development set contains 5870 audio files from 658 speakers with a minimum of 8 sessions/channels per speaker. For enrollment, we predefine 6 source files for each speaker, thus providing 6 × 8 retransmitted segments that we could choose from. Accordingly, three different test conditions can be identified as: (i) seen, (ii) unseen, and (iii) both. Speaker models for the seen and unseen conditions are enrolled using numerous channels selected randomly for each of the 6 enrollment segments. A test is labeled as "seen" trial if the model against which it is scored has observed the test channel during enrollment. If the speaker model has not observed the test channel during enrollment, the associated trial is part of "unseen" trials. The "both" condition is the combination of seen and unseen trial subsets. The total number of trials for each of the 3 test conditions are: seen (1,338,811), unseen (1,092,956), and both (2,431,767).

#### 3.2.2. LID task

A total of 16,000 speech recordings, including clean source files along with their degraded (retransmitted) versions (channels A through H), from the data releases LDC2011E95, LDC2011E111, LDC2012E06 for the RATS LID task (RATS, 2013), are partitioned into development, enrollment, and test sets. The recordings are extracted from found corpora including the NIST LRE data, CallFriend, and Fisher. The development set, which is formed by pooling enrollment data from all languages, contains 15,000 audio files from five target languages including Levantine Arabic (3256), Dari (854), Farsi (1037), Pashto (2947), and Urdu (3239) as well as 10 impostor languages (labeled as xxx) including English, Spanish, Mandarin, Thai, Vietnamese, Russian, Japanese, Bengali, Korean, and Tagalog. The test set contains 1000 audio files from the five target as well as 10 impostor languages. The total number of recordings for each of the language categories in the test set is as follows: Levantine Arabic (152), Dari (38), Farsi (189), Pashto (122), Urdu (143), and xxx (356).

### 3.3. System description

For both SID and LID experiments, non-speech frame dropping is performed on the acoustic feature vectors for each of the different front-ends considered in this study. We use speech/non-speech labels generated with a

Table 1
Configuration parameters for the MHEC as well as other front-ends considered in our study.

| Frontend | Configuration parameter | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Frame length (ms) | Frame shift (ms) | Filterbank | No. bands | Freq. scale | No. coeffs. | $c_0$ | Post-process |
| MFCC | 25 | 10 | triangular | 32 | mel | $20 + \Delta + \Delta\Delta$ | Yes | CMVN |
| MHEC | 25 | 10 | Gammatone | 32 | ERB | $20 + \Delta + \Delta\Delta$ | Yes | CMVN |
| PNCC | 25 | 10 | Gammatone | 32 | ERB | $20 + \Delta + \Delta\Delta$ | Yes | CMVN |
| PMVDR | 25 | 10 | – | – | mel | $20 + \Delta + \Delta\Delta$ | Yes | CMVN |

supervised speech activity detector (SAD) (Graciarena et al., 2013) for our SID experiments, while for our LID evaluations, time labels generated with the unsupervised Combo-SAD (Sadjadi and Hansen, 2013b) are employed. After dropping the non-speech frames, global (utterance level) cepstral mean and variance normalization (CMVN) is applied to suppress the short-term linear channel effects.

### 3.3.1. SID system

We perform our SID experiments in the context of a state-of-the-art i-vector based system. To learn the i-vector extractor (Dehak et al., 2011), a gender-independent 512-component universal background model (UBM) with diagonal covariance matrices is first trained in a binary split Expectation–Maximization (EM) framework using a total of 5870 900-s long recordings from 658 speakers in the development set. The zeroth and first order Baum-Welch statistics are then computed for each recording and used to learn a 400-dimensional total variability subspace. After extracting 400-dimensional i-vectors, we use linear discriminant analysis (LDA) to reduce the dimensionality to 200. The dimensionality reduced i-vectors are then centered (the mean is removed) and length normalized. For scoring, a Gaussian probabilistic LDA (PLDA) model with full covariance residual noise term (Prince and Elder, 2007; Garcia-Romero and Espy-Wilson, 2011) is learned using the i-vectors extracted from 5870 30-s cuts in the development set. The Eigenvoice matrix in the PLDA model is full-rank with 200 columns. Given the size of the experiments and the number of front-ends to be evaluated, in this study we only consider the 30–30 s evaluation condition from the RATS SID task. Audio segments for enrollment and test in the 30–30 s condition are 30-s cuts extracted from full duration 900-s recordings. Here, the simple yet effective i-vector averaging strategy is employed for 6-sided speaker enrollment, yielding a total of 2936 speaker models. There are 1061 30-s test segments for the evaluation.

### 3.3.2. LID system

Our LID experiments are conducted in the context of a Joint Factor Analysis (JFA) system (Kenny et al., 2007). A 1024-component language-independent UBM with diagonal covariance matrices is trained with the EM algorithm using 15,000 150-s long (on average) recordings in the development set. A JFA framework without Eigenvoices is employed to model the languages. The $\boldsymbol{D}$ matrix is initialized with maximum a posteriori (MAP) adaptation from the UBM with a relevance factor of $\tau = 0.1$. A 100-dimensional channel matrix $\boldsymbol{U}$ is used along with a linear scoring strategy to obtain final scores for each test segment (5 for the target and one for the impostor languages). Here, similar to our SID experiments, we only consider the 30sec evaluation condition from the RATS LID task, although the language models are trained using full duration audio recordings.

## 4. Results and discussion

### 4.1. SID experiments

Table 2 presents results obtained from our experiments on different noisy and channel degraded evaluation conditions available in the RATS SID task described previously. The results are reported in terms of EER as well as false-alarm rate at a 10% false-rejection rate (FA10m), which is the RATS program performance metric for Phase I. The EER is defined as a point on the detection error trade-off (DET) curve where the probability of false-rejects ($P_{\text{miss}}$) is equal to that of false-alarms ($P_{\text{fa}}$). Here, we evaluate the effectiveness of acoustic features extracted with the MHEC front-end against those generated with MFCC, PMVDR, and PNCC front-ends. In addition, SID results are shown for MHEC and PNCC features with both log and root compression strategies. Several observations can be made from the results given in Table 2.

First, irrespective of the nonlinear operator used for dynamic range compression, the MHEC feature provides better SID performance, in terms of both EER and FA10m metrics, compared to the baseline MFCCs and the alternative noise-robust front-ends across different evaluation conditions. The effectiveness of the MHEC is particularly more evident under the challenging “unseen” condition where speaker models have not been exposed to the channel condition in the tests.

Second, for both the MHEC and PNCC features, significantly better SID performance is achieved with the power-law (plaw) nonlinearity compared to the traditional logarithmic compression. As noted earlier, there are two major reasons for this behavior; (1) the root compression with an exponent of $\gamma = 1/15$ approximates more accurately (compared to the log compression) the nonlinear relation between the input sound intensity and auditory nerve firing rate, and (2) the log compression artificially increases the variance for spectral components with values much smaller than one, resulting in a greater variability of the estimated cepstral features. As we elaborated before, this increased variability may not remain local for spectral energies computed for a specific channel and is spread across all cpestral coefficients due to the DCT operation.

### 4.2. LID experiments

Results obtained from our LID experiments on the individual languages and front-ends consiRATS LID task are shown in Figs. 6 and 7. The results are reported in terms of EER and Cavg (NIST, 2009) for dered in this study. The Cavg is defined as follows:

$$C_{avg} = \frac{1}{N_L}\sum_{L_T}\left\{\begin{array}{ll} & C_{miss}\cdot P_{\text{Target}}\cdot P_{miss}(L_T) \\ + & \displaystyle\sum_{L_N} C_{fa}\cdot P_{\text{NonTarget}}\cdot P_{fa}(L_T, L_N) \\ + & C_{fa}\cdot P_{\text{Out-of-set}}\cdot P_{fa}(L_T, L_O) \end{array}\right\},$$

Please cite this article in press as: Sadjadi, S.O., Hansen, J.H.L., Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.04.005

Table 2
Performance comparison of alternative noise-robust features considered in this study against MFCCs on the RATS SID task under seen and unseen 30–30 s enrollment/test conditions as well as their combination (both), in terms of percent EER and FA10m (in parentheses). "plaw" denotes the power law compression.

| Condition | EER (FA10m) (%) | | | | | |
|---|---|---|---|---|---|---|
| | MFCC | PMVDR | MHEC-log | MHEC-plaw | PNCC-log | PNCC-plaw |
| Seen | 8.37<br>(6.85) | 8.34<br>(6.75) | 7.44<br>(5.31) | **7.09**<br>**(4.62)** | 7.76<br>(5.78) | 7.34<br>(5.29) |
| Unseen | 8.68<br>(7.49) | 8.65<br>(7.46) | 7.52<br>(5.61) | **7.16**<br>**(4.95)** | 7.86<br>(6.03) | 7.63<br>(5.68) |
| Both | 8.52<br>(7.18) | 8.50<br>(7.08) | 7.49<br>(5.48) | **7.14**<br>**(4.80)** | 7.81<br>(5.91) | 7.46<br>(5.48) |
| Avg. | 8.52<br>(7.17) | 8.50<br>(7.10) | 7.48<br>(5.46) | **7.13**<br>**(4.79)** | 7.81<br>(5.91) | 7.48<br>(5.48) |

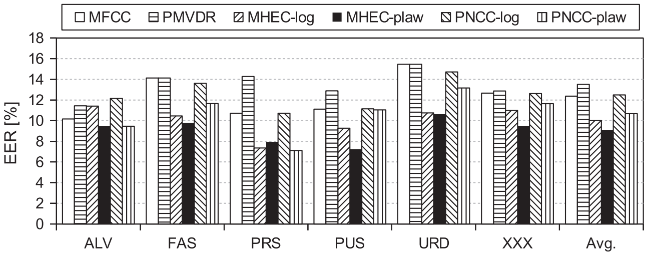

Fig. 6. Performance comparison of alternative noise-robust features considered in this study against MFCCs on the RATS LID task (30 s test condition) in terms of percent EER. XXX denotes all impostor languages.

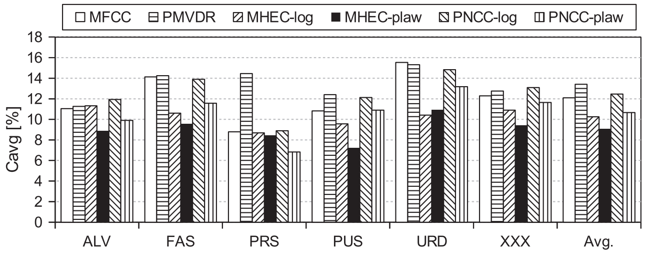

Fig. 7. Performance comparison of alternative noise-robust features considered in this study against MFCCs on the RATS LID task (30 s test condition) in terms of percent Cavg. XXX denotes all impostor languages.

where $L_T, L_N$, and $L_O$ are target, non-target, and out-of-set languages, and $C_{miss} = 1, C_{fa} = 1$, and $P_{\text{Target}} = 0.5$ are application model parameters. Here, $N_L$ is the number of closed-set languages. The parameters $P_{\text{Out-of-set}}$ and $P_{\text{NonTarget}}$ are set as,

$$P_{\text{Out-of-set}} = \begin{cases} 0.0 & \text{for closed-set conditions} \\ 0.2 & \text{for open-set conditions} \end{cases},$$

and

$$P_{\text{NonTarget}} = \frac{(1 - P_{\text{Target}} - P_{\text{Out-of-set}})}{N_L - 1}.$$

In a similar trend observed in the RATS SID experiments, it can be seen from the figure that the MHEC feature consistently provides significant LID performance improvements over the baseline MFCCs across all five target languages, i.e., Levantine Arabic (alv), Dari (prs), Farsi (fas), Pashto (pus), and Urdu (urd), as well as the 10 out-of-set languages (xxx). This is irrespective of the nonlinear operator used in the dynamic range compression stage. In addition, on average (avg) and for almost all language classes, the MHEC outperforms the alternative noise-robust front-ends, in terms of both EER and Cavg metrics.

Also shown in Figs. 6 and 7 are results obtained from our LID experiments with the MHEC and PNCC features using both log and root compression strategies. Consistent with the results obtained from the SID experiments, it is clear from the figure that for both front-ends there are significant improvements in LID performance with the power-law nonlinearity. The improvements are observed on average (avg) and across almost all languages considered, in terms of EER and Cavg.

It is worth remarking here that, although in this study we only considered the 30-s test condition from the RATS tasks to evaluate the effectiveness of the proposed acoustic front-end for SID and LID, it has been shown in several independent studies (McLaren et al., 2013; Mitra et al., 2013; Lawson et al., 2013; Matejka et al., 2014) that the MHEC feature provides similar levels of improvement against MFCCs for a range of test duration conditions (i.e., 3 s, 10 s, 30 s, 120 s), in particular for shorter duration tasks.

## 5. Conclusion

Significant advancements have been made in recent years towards more effective SID and LID systems, yet adverse noisy mismatch conditions can cause major drops in performance. The present study has proposed a robust alternative acoustic feature representation, entitled the mean Hilbert envelope coefficients (MHEC), to overcome the limitations of the conventional MFCCs which are the most widely used features in many speech applications. The proposed acoustic feature was based on the Hilbert transform demodulation of subband signals obtained from

Please cite this article in press as: Sadjadi, S.O., Hansen, J.H.L., Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.04.005

the Gammatone filterbank analysis of speech. The subband Hilbert envelopes represent the envelope of the auditory nerve response at specific center frequencies. They carry useful acoustic information and have been shown to be robust to signal degradations. Through a set of experimental evaluations in the context of noisy and channel degraded RATS SID and LID tasks, the effectiveness of the MHEC feature was confirmed and it was shown to outperform the baseline MFCCs as well as PMVDR and PNCC front-ends, in particular under the more challenging mismatch ("unseen") conditions. In addition, we investigated the impact of the dynamic range compression stage on SID and LID performance for both MHEC and PNCC features. The power-law nonlinearity was observed to yield the best performance across different conditions for both SID and LID tasks.

## References

Alam, M.J., Kinnunen, T., Kenny, P., Ouellet, P., O'Shaughnessy, D., 2013. Multitaper MFCC and PLP features for speaker verification using i-vectors. Speech Commun. 55 (2), 237–251.

Alexandre, P., Lockwood, P., 1993. Root cepstral analysis: a unified view. Application to speech processing in car noise environments. Speech Commun. 12 (3), 277–288.

Assmann, P., Summerfield, A., 2004. The perception of speech under adverse conditions. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (Eds.), Speech Processing in the Auditory System. Springer-Verlag, New York.

Bielefeld, B., 1994. Language identification using shifted delta cepstrum. In: Proc. 14th Annual Speech Research Symposium.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics Speech Signal Process. 27 (2), 113–120.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C., 2007. Compensation of nuisance factors for speaker and language recognition. IEEE Trans. Audio Speech Lang. Process. 15 (7), 1969–1978.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95 (5), 2670–2680.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, 2nd Edition. Academic Press (Chapter 10).

Ganapathy, S., Thomas, S., Hermansky, H., 2012. Feature extraction using 2-D autoregressive models for speaker recognition. In: Proc. ISCA Odyssey. Singapore, Singapore.

Garcia-Romero, D., Espy-Wilson, C., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Proc. INTERSPEECH. Florence, Italy, pp. 249–252.

Glasberg, B.R., Moore, B.C., 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Res. 47 (12), 103–138.

Godin, K.W., Sadjadi, S.O., Hansen, J.H.L., 2013. Impact of noise reduction and spectrum estimation on noise robust speaker identification. In: Proc. INTERSPEECH. Lyon, France, pp. 3656–3660.

Graciarena, M., Alwan, A., Ellis, D., Franco, H., Ferrer, L., Hansen, J.H.L., Janin, A., Lee, B.-S., Lei, Y., Mitra, V., Morgan, N., Sadjadi, S.O., Tsai, T., Scheffer, N., Tan, L.N., Williams, B., 2013. All for one: feature combination for highly channel-degraded speech activity detection. In: Proc. INTERSPEECH. Lyon, France, pp. 709–713.

Han, K.J., Ganapathy, S., Li, M., Omar, M.K., Narayanan, S., 2013. Trap language identification system for RATS phase II evaluation. In: Proc. INTERSPEECH. Lyon, France, pp. 1502–1506.

Hanilci, C., Kinnunen, T., Ertas, F., Saeidi, R., Pohjalainen, J., Alku, P., 2012a. Regularized all-pole models for speaker verification under noisy environments. IEEE Signal Process. Lett. 19, 163–166.

Hanilci, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., Ertas, F., Sandberg, J., Hansson-Sandsten, M., 2012b. Comparing spectrum estimators in speaker verification under additive noise degradation. In: Proc. IEEE ICASSP. Kyoto, Japan, p. 47694772.

Hasan, T., Sadjadi, S.O., Liu, G., Shokouhi, N., Bořil, H., Hansen, J.H.L., 2013. CRSS systems for 2012 NIST speaker recognition evaluation. In: Proc. IEEE ICASSP. Vancouver, BC, pp. 6783–6787.

Hatch, A., Kajarekar, S., Stolcke, A., 2006. Within-class covariance normalization for SVM-based speaker recognition. In: Proc. INTERSPEECH. Pittsburgh, PA, pp. 1471–1474.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87 (4), 1738–1752.

Hu, G., Wang, D., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks 15 (5), 1135–1150.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus Eigenchannels in speaker recognition. IEEE Trans. Audio Speech Lang. Process. 15 (4), 1435–1447.

Kim, C., Stern, R.M., 2012. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: Proc. IEEE ICASSP. Kyoto, Japan, pp. 4101–4104.

Kingsbury, B.E., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Commun. 25 (13), 117–132.

Kinnunen, T., Rajan, P., 2013. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In: Proc. IEEE ICASSP. Vancouver, BC, pp. 7229–7233.

Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K.A., Sandberg, J., Hansson-Sandsten, M., Li, H., 2012. Low-variance multitaper MFCC features: a case study in robust speaker verification. IEEE Trans. Audio Speech Lang. Process. 20 (7), 1990–2001.

Kinnunen, T., Alam, M.J., Matějka, P., Kenny, P., Černocky, J., OShaughnessy, D., 2013. Frequency warping and robust speaker verification: a comparison of alternative mel-scale representations. In: Proc. INTERSPEECH. Lyon, France, pp. 3122–3126.

Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., Stauffer, A., 2011. Survey and evaluation of acoustic features for speaker recognition. In: Proc. IEEE ICASSP. Prague, Czech Republic, pp. 5444–5447.

Lawson, A., McLaren, M., Lei, Y., Mitra, V., Scheffer, N., Ferrer, L., Graciarena, M., 2013. Improving language identification robustness to highly channel-degraded speech through multiple system fusion. In: Proc. INTERSPEECH. Lyon, France, pp. 1507–1510.

Li, Q., Huang, Y., 2010. Robust speaker identification using an auditory-based feature. In: Proc. IEEE ICASSP. Dallas, TX, pp. 4514–4517.

Lim, J., 1979. Spectral root homomorphic deconvolution system. IEEE Trans. Acoust. Speech Signal Process. 27 (3), 223–233.

Maragos, P., Kaiser, J.F., Quatieri, T.F., 1993. Energy separation in signal modulations with application to speech analysis. IEEE Trans. Signal Process. 41 (10), 3024–3051.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512.

Matějka, P., Plchot, O., Soufifar, M., Glembek, O., D'Haro, L., Veselý, K., Grézl, F., Ma, J., Matsoukas, S., Dehak, N., 2012. Patrol team language identication system for DARPA RATS P1 evaluation. In: Proc. INTERSPEECH. Portland, OR.

Matejka, P., Zhang, L., Ng, T., Mallidi, S.H., Glembek, O., Ma, J., Zhang, B., 2014. Neural network bottleneck features for language identification. In: Proc. Odyssey 2014: The Speaker and Language Recognition Workshop. Joensuu, Finland, pp. 299–304.

Please cite this article in press as: Sadjadi, S.O., Hansen, J.H.L., Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Comm. (2015), http://dx.doi.org/10.1016/j.specom.2015.04.005

McLaren, M., Scheffer, N., Graciarena, M., Ferrer, L., Lei, Y., 2013. Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion. In: Proc. IEEE ICASSP. Vancouver, BC, pp. 6773–6777.

Mitra, V., McLaren, M., Franco, H., Graciarena, M., Scheffer, N., 2013. Modulation features for noise robust speaker identification. In: Proc. INTERSPEECH. Lyon, France, pp. 3703–3707.

NIST, 2009. The NIST Year 2009 Language Recognition Evaluation (LRE) Plan. <http://www.itl.nist.gov/iad/mig/tests/lre/2009/>.

Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M., 1992. Complex sounds and auditory images. In: Cazals, Y., Demany, L., Horner, K. (Eds.), Auditory Physiology and Perception. Pergamon Press, Oxford, pp. 429–446.

Prince, S., Elder, J., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: Proc. IEEE Int. Conf. Computer Vision, ICCV 2007. Rio de Janeiro, pp. 1–8.

RATS, 2013. DARPA Robust Automatic Transcription of Speech (RATS). <http://projects.ldc.upenn.edu/RATS/>.

Ravindran, S., Anderson, D.V., Slaney, M., 2006. Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing. In: Proc. ISCA SAPA. Pittsburgh, PA, pp. 48–52.

Sadjadi, S.O., Hansen, J.H.L., 2010. Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. In: Proc. INTERSPEECH. Makuhari, Japan, pp. 2138–2141.

Sadjadi, S.O., Hansen, J.H.L., 2011. Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In: Proc. IEEE ICASSP. Prague, Czech Republic, pp. 5448–5451.

Sadjadi, S.O., Hansen, J.H.L., 2012. Blind reverberation mitigation for robust speaker identification. In: Proc. IEEE ICASSP. Kyoto, Japan, pp. 4225–4228.

Sadjadi, S.O., Hansen, J.H.L., 2013a. Robust front-end processing for speaker identification over extremely degraded communication channels. In: Proc. IEEE ICASSP. Vancouver, BC, pp. 7214–7218.

Sadjadi, S.O., Hansen, J.H.L., 2013b. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process. Lett. 20, 197–200.

Sadjadi, S.O., Hansen, J.H.L., 2014. Blind spectral weighting for robust speaker identification under reverberation mismatch. IEEE Trans. Audio Speech Lang. Process. 22 (5), 935–943.

Sadjadi, S.O., Hasan, T., Hansen, J.H.L., 2012. Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition. In: Proc. INTERSPEECH. Portland, OR, pp. 1696–1699.

Sarikaya, R., Hansen, J.H.L., 2001. Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition. In: Proc. INTERSPEECH. Aalborg, Denmark, pp. 687–690.

Schimmel, S.M., 2007. Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices. Ph.D. Thesis, Dept. Elect. Eng., University of Washington.

Shao, Y., Srinivasan, S., Wang, D., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: Proc. IEEE ICASSP. Honolulu, HI, pp. 277–280.

Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller, J.R., 2002. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In: Proc. INTERSPEECH. Denver, CO, pp. 33–36.

Vakman, D., 1994. On the analytic signal, the Teager–Kaiser energy algorithm, and other methods for defining amplitude and frequency. IEEE Trans. Signal Process. 44 (4), 791–797.

Walker, K., Strassel, S., 2012. The RATS radio traffic collection system. In: Proc. ISCA Odyssey. Singapore, Singapore.

Weintraub, M., 1985. A Theory and Computational Model of Auditory Monaural Sound Separation. Ph.D. Thesis, Dept. Elect. Eng., Stanford University.

Yapanel, U., Hansen, J.H.L., 2008a. A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition. Speech Commun. 50, 142–152.

Yapanel, U., Hansen, J.H.L., 2008b. Towards an intelligent acoustic front end for automatic speech recognition: built-in speaker normalization. Eurasip J. Audio Speech Music Process. 2008, 1–13.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2009. HTK – Hidden Markov Model Toolkit v3.4.1. <http://htk.eng.cam.ac.uk/>.

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., Shamma, S., 2011. Linear versus mel frequency cepstral coefficients for speaker recognition. In: Proc. IEEE ASRU. Hawaii, HI, pp. 559–564.